УДК 004.8:338+629.33

А. В. Лосенко

РОЗРОБКА ІНФОРМАЦІЙНОЇ СИСТЕМИ ПЕРЕДБАЧЕННЯ ЦІНИ НА ВЖИВАНІ АВТО

Вінницький національний технічний університет

Анотація

Був здійснений розвідувальний аналіз обраного набору даних, розміщеного у вільному доступі, за допомогою інструментів аналізу даних написаних на мові програмування Python, після чого була обрана найбільш ефективна модель машинного навчання для вирішення завдання передбачення ціни вживаного автомобіля.

Ключові слова: *розвідувальний аналіз даних, передбачення ціни, вживаний автомобіль*

Abstract

An exploratory data analysis of the selected set of open dataset was concluded using data analysis tools written in Python programming language, and then the most efficient machine learning model was selected to solve the problem of predicting the price of a used car.

Keywords: *exploratory data analysis, price forecasting, used car*

Вступ

На ринку вживаних автомобілів в Україні спостерігається щорічний ріст, що свідчить про потребу в інструментах, які допоможуть прийняти рішення під час формування ціни на той чи інший автомобіль, базуючись на його властивостях. За допомогою такого інструменту продавець автомобілю буде здатний сформувати найвигіднішу ціну, що в свою чергу допоможе йому ефективніше виконувати свою роботу та приносити більші прибутки власній справі. Отже, в такому випадку, розробка засобу, що служитиме в описаній ситуації є більш ніж актуальним завданням.

Метою роботи є застосування методів розвідувального аналізу та машинного навчання для вирішення задачі передбачення ціни вживаного автомобіля.

Постановка задачі

Для дослідження був обраний набір даних, що розміщений у вільному доступі на веб-платформі Kaggle [1]. Даний датасет містить в собі дані про вживані автомобілі з веб-сайту craigslist.com. Спочатку був проведений розвідувальний аналіз даних (exploratory data analysis, EDA) [2], з метою загального огляду вмісту набору набору даних, його очистки, а також визначення можливих закономірностей між його ознаками. Для обробки та аналізу даних були обрані програмні пакети мови програмування Python, серед яких були застосовані бібліотеки Pandas, Numpy, Scikit-Learn, та інші [3]. Загальна кількість елементів датасету дорівнює 509577, серед яких 355200 елементів з відсутніми значеннями у деяких ознак. У зв'язку з цим датасет був очищений від даних з неповним обсягом інформації, що зменшило його до розміру в 154377 записів. Далі набір даних був відфільтрований за ознаками, які в подальшому будуть використані при тренуванні моделей машинного навчання, такими ознаками є:

- модель автомобіля (“make”);
- виробник автомобіля (“manufacturer”);
- стан автомобіля (“condition”);
- рік випуску (“year”);
- вид палива (“fuel”), набуває одне з двох значень (“gas”, “diesel”) ;
- пробіг автомобіля (“odometer”), вимірюється у км. (25000, 120000, 30000...);
- трансмісія (“transmission”), набуває одне з двох значень (“automatic”, “manual”);
- привід автомобіля (“drive”), набуває одне з трьох значень (“fwd”, “rwd”, “4wd”) ;

тип кузову (“type”) набуває одне з 5 значень (“coupe”, “SUV”, “wagon”, “minivan”, “pickup”);

Приклад елементів датасету наведений на рисунку 1:

price	year	manufacturer	model	condition	cylinders	fuel	odometer	transmission	drive	type
17899	2012.0	volkswagen	golf r	excellent	4 cylinders	gas	63500.0	manual	4wd	hatchback
4600	2008.0	honda	civic	good	4 cylinders	gas	110982.0	automatic	fwd	sedan
28000	2004.0	ford	f550 mechanics service	good	10 cylinders	gas	67348.0	automatic	4wd	truck
18999	2015.0	mercedes-benz	cla-class	excellent	4 cylinders	gas	37000.0	automatic	fwd	sedan
79997	2016.0	mercedes-benz	amg gts	excellent	8 cylinders	gas	28000.0	automatic	rwd	coupe

Рис. 1. Приклад даних з датасету вживаних автомобілів

Після очистки та фільтрування даних датасет був також зменшений у розмірі, за рахунок заміни типів даних, у яких виражені ознаки. Крім цього, були відфільтровані аномальні дані, що виходили за межі значень в квантилях 10% та 90%, за рахунок цього був отриманий набір даних з найбільш актуальними значеннями. Всі ці заходи посприяли швидшій обробці даних на етапі тренування моделями машинного навчання. Програмний код, що містить процедури розвідувального аналізу даних обраного датасету доступний на кернелі веб-платформи Kaggle [4].

Результати дослідження

Для порівняння ефективності моделей в задачі передбачення ціни вживаного автомобіля були обрані 15 моделей, кожна з яких була натренована на відфільтрованому датасеті, після чого результати передбачень були порівняні за кількома критеріями та методами, серед яких найвагомішим був коефіцієнт детермінації R^2.

Даний коефіцієнт є показником залежності варіації залежної змінної від варіації незалежних змінних, що свідчать наскільки модель підтверджується значеннями отриманих спостережень

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (1)$$

де y — істинне значення з тренувального набору даних; $\hat{y}$— передбачене значення відповідного набору ознак; $\bar{y}$— середнє арифметичне істинних значень з тренувального масиву даних.

Спираючись на значення даного коефіцієнту при передбаченні ціни вживаних автомобілів з тестової вибірки даних були виділені 5 найефективніших моделей, серед яких були наступні:

- LGBM (LightGBM);
- Extra Trees Regressor;
- Random Forest;
- Bagging Regressor;
- Gradient Boosting Regressor;

Також для оцінювання ефективності моделей доцільно використати величини RMSE (root-mean-square-error, середньоквадратична похибка) та відносної похибки δ на основі вбудованої функції MAE (абсолютне значення середньої похибки)[5].

Результати тестування точності передбачення даних моделей зображені на рисунку 2:

Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
LGBM	91.09	85.62	12.36	14.34	178,664.68	210,863.23
ExtraTreesRegressor	99.97	83.58	0.09	13.50	10,492.81	230,648.45
Random Forest	97.26	83.35	5.65	14.27	95,416.65	228,713.74
BaggingRegressor	97.15	83.29	5.74	14.35	97,230.64	229,744.42
GradientBoostingRegressor	85.30	82.77	14.84	15.80	210,905.93	228,370.36

Рис. 2. Приклад даних з датасету вживаних автомобілів

Як видно з наведеного вище рисунку, найефективнішою моделлю машинного навчання для задачі передбачення ціни вживаного авто є модель LGBM (LightGBM) [6].

Висновки

За допомогою обраних програмних пакетів мови програмування Python було досліджено набір даних продажів вживаних авто у США. Під час дослідження був здійснений розвідувальний аналіз даних, що дав змогу відфільтрувати неактуальні ознаки, а також зменшити об'єм масиву даних. Після попереднього аналізу відфільтрований датасет був оброблений моделями машинного навчання, після чого на основі значення коефіцієнту детермінації була обрана найбільш оптимальна модель для поставленої задачі, а саме модель LGBM з бібліотеки lightgbm, з точністю 85,62%.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Used Cars Dataset, Vehicles listtings from Craigslist [Electronic resource] Available: https://www.kaggle.com/austinreese/craigslist-carstrucks-data

2. Jean-Daniel Fekete, Romain Primet, "Progressive Analytics: A Computation Paradigm for Exploratory Data Analysis", https://arxiv.org/abs/1607.05162

3. Wes McKinney "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd. ed.).", *O'Reilly Media, Inc, 2017*.

4. Used Cars - FE & EDA with 3D, abnormals filter [Electronic resource] Available: https://www.kaggle.com/vbmokin/used-cars-fe-eda-with-3d-abnormals-filter

5. Мокін В. Б., Лосенко А. В., Дратований М. В., Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі [Електронний ресурс] Режим доступу: https://doi.org/10.31649/1997-9266-2019-147-6-62-72

6. LightGBM Documentation [Electronic resource] Available: https://lightgbm.readthedocs.io/en/latest/

Лосенко Арсен Володимирович — аспірант кафедри системного аналізу, комп'ютерного моніторингу та інженерної графіки, e-mail: arsenlosenko@protonmail.com

Науковий керівник: ***Мокін Віталій Борисович*** — д-р техн. наук, професор, завідувач кафедри системного аналізу, комп'ютерного моніторингу та комп'ютерної графіки.

Losenko Arsen Volodymyrovych – Post-Graduate Student of the Department of System Analysis, Computer Monitoring and Engineering Graphics, e-mail: arsenlosenko@protonmail .com

Supervisor: ***Mokin Vitalii Borysovych*** - Dr. Tech. Sciences, Professor, Head of the Department of System Analysis, Computer Monitoring and Computer Graphics.